WHAT IS CLAIMED IS:

Claims

1. A method for annotating a query email message, the method comprising the steps of:

accessing patterns associated with a database comprising annotated email messages;

assigning attributes to the patterns based on the annotated email messages; and

using the patterns with assigned attributes to analyze the query email message.

2. The method of Claim 1, wherein the step of accessing patterns comprises using a pattern discovery algorithm.

3. The method of Claim 1, wherein the pattern discovery algorithm is the Teiresias pattern algorithm.

4. The method of Claim 1, wherein the steps of accessing patterns and assigning attributes are carried out independently of and prior to the step of using the patterns with assigned attributes to analyze the query email message.

5. The method of Claim 1, further comprising the step of selecting the accessed patterns that match the query email message.

6. The method of claim 1, further comprising the step of storing the patterns with with assigned attributes in a database.

7. The method of claim 1, wherein the using step further comprises the step of defining an attribute vector from the patterns with assigned attributes, the attribute vector characterizing portions of the query email message.

8. The method of claim 1, wherein the using step further comprises the step of defining an attribute vector from the patterns with assigned attributes, the attribute vector characterizing the whole of the query email message.

9. The method of claim 1, wherein one or more of said annotated email messages comprises an unwelcome email message ("SPAM").

10. The method of claim 9, further comprising the step of storing the patterns with assigned attributes in a database serving as a "SPAM-dictionary".

11. The method of claim 1, wherein one or more of said annotated email messages comprises a welcome email message ("non-SPAM").

12. The method of claim 11, further comprising the step of storing the patterns with assigned attributes in a database serving as a "SPAM-dictionary".

13. The method of claim 1, wherein said database comprises (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM").

14. The method of claim 7, wherein the attribute vector comprises a number of counters.

15. The method of claim 14, wherein the query email message comprises characters of a human language and the number of counters is proportional to the number of said characters in the query email message.

16. The method of claim 14, wherein the assigned attributes are used to contribute values to counters of the attribute vector corresponding to portions of the query email message matched by the patterns.

17 The method of claim 7, comprising a plurality of attribute vectors.

18. The method of claim 17, wherein the values contributed to the counters of each of the attribute vectors of the plurality of attribute vectors are normalized.

19. The method of claim 17, wherein each attribute vector of the plurality of attribute vectors represents a different attribute.

20. The method of claim 17, wherein the plurality of attribute vectors are ranked.

21. The method of claim 20, wherein only highly ranking attribute vectors are kept.

22. The method of claim 1, further comprising the step of determining a score for the patterns with assigned attributes used to contribute to the attribute vector.

23. The method of claim 22, wherein the score represents a degree of similarity between the query email message and at least one annotated email message of the database.

24. The method of claim 23, wherein the score is normalized.

25. The method of claim 22, wherein the score represents a degree of similarity between the query email message and at least one annotated email message of the database, and wherein said at least one of said annotated email messages comprises an unwelcome email message ("SPAM").

26. The method of claim 22, wherein the score represents a degree of similarity between the query email message and at least one annotated email message of the database, and wherein said at least one of said annotated email messages comprises a welcome email message ("non-SPAM").

27. The method of claim 1, further comprising the step of determining a score for the patterns with assigned attributes used to contribute to the attribute vector, said database comprising (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM"), said score representing a degree of similarity, between the query email message and at least one of said annotated unwelcome email messages ("SPAM"), and a degree of dissimilarity between the query email message and at least one of said annotated welcome email messages ("non-SPAM").

28. The method of claim 27, further comprising the step of defining, for each of said assigned attributes, a value criterion based on the value of the counters of the attribute vector to determine whether the corresponding attribute is present in the query email message.

29. The method of claim 27, further including the step of defining a SPAM attribute criterion dependent on which of said assigned attributes are present in the query email message, to determine whether the query email message is a SPAM email message.

30. The method of claim 27, further including the step of defining a non-SPAM attribute criterion dependent on which of said assigned attributes are present in the query email message, to determine whether the query email message is a non-SPAM email message

31. An apparatus for annotating a query email message, the apparatus comprising:

a memory; and

at least one processor, coupled to the memory, operative to:

access patterns associated with a database comprising annotated email messages;

assign attributes to the patterns based on the annotated email messages; and

use the patterns with assigned attributes to analyze the query email message.

32. The apparatus of claim 31, wherein the at least one processor is further operative to select the accessed patterns that match the query email message.

33. The apparatus of claim 31, wherein in accordance with the using operation the at least one processor is further operative to define an attribute vector from the patterns with assigned attributes, the attribute vector characterizing portions of the query email message.

34. The apparatus of claim 31, wherein at least one of said annotated email messages comprises an unwelcome email message ("SPAM").

35. The apparatus of claim 31, wherein at least one of said annotated email messages comprises a welcome email message ("non-SPAM").

36. The apparatus of claim 31, wherein said database comprises (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM").

37. The apparatus of claim 33, wherein the attribute vector comprises a number of-counters.

38. The apparatus of claim 37, wherein the query email message comprises characters of a human language and the number of counters is proportional to the number of said characters in the query email message.

39. The apparatus of claim 37, wherein the assigned attributes are used to contribute values to counters of the attribute vector corresponding to portions of the query email message matched by the patterns.

40 The apparatus of claim 33, comprising a plurality of attribute vectors.

41. The apparatus of claim 39, wherein each attribute vector of the plurality of attribute vectors represents a different attribute.

42. The apparatus of claim 39, wherein the plurality of attribute vectors are ranked.

43. The apparatus of claim 31, wherein the at least one processor is further operative to determine a score for the patterns with assigned attributes used to contribute to the attribute vector.

44. The apparatus of claim 43, wherein the score represents a degree of similarity between the query email message and the annotated email messages of the database.

45. The apparatus of claim 43, wherein the score represents a degree of similarity between the query email message and at least one of the annotated email messages of the database, and wherein said at least one of said annotated email messages comprises an unwelcome email message ("SPAM").

46. The apparatus of claim 43, wherein the score represents a degree of similarity between the query email message and at least one of the annotated email messages of the database, and wherein said at least one of said annotated email messages comprises a welcome email message ("non-SPAM").

47. The apparatus of claim 31, wherein the at least one processor is further operative to determine a score for the patterns with assigned attributes used to contribute to the attribute vector, said database comprising (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM"), said score representing a degree of similarity, between the query email message and said annotated unwelcome email messages ("SPAM"), and a degree of dissimilarity between the query email message and said annotated welcome email messages ("non-SPAM").

48. An article of manufacture for annotating a query email message, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

accessing patterns associated with a database comprising annotated email messages;

assigning attributes to the patterns based on the annotated email messages; and

using the patterns with assigned attributes to analyze the query email message.

49. The article of manufacture of claim 48, further comprising the step of selecting the accessed patterns that match the query email message.

50. The article of manufacture of claim 48, wherein the using step further comprises defining an attribute vector from the patterns with assigned attributes, the attribute vector characterizing portions of the query email message.

51. The article of manufacture of claim 48, wherein at least one of said annotated email messages comprises an unwelcome email message ("SPAM").

52. The article of manufacture of claim 48, wherein at least one of said annotated email messages comprises a welcome email message ("non-SPAM").

53. The article of manufacture of claim 48, wherein said database comprises (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM").

54. The article of manufacture of claim 50, wherein the attribute vector comprises a number of-counters.

55. The article of manufacture of claim 54, wherein the query email message comprises characters in a human language and the number of counters is proportional to the number of said characters in the query email message.

56. The article of manufacture of claim 54, wherein the assigned attributes are used to contribute values to counters of the attribute vector corresponding to portions of the query email message matched by the patterns.

57 The article of manufacture of claim 50, comprising a plurality of attribute vectors.

58. The article of manufacture of claim 57, wherein each attribute vector of the plurality of attribute vectors represents a different attribute.

59. The article of manufacture of claim 57, wherein the plurality of attribute vectors are ranked.

60. The method of claim 48, further comprising the step of determining a score for the patterns with assigned attributes used to contribute to the attribute vector.

61. The article of manufacture of claim 60, wherein the score represents a degree of similarity between the query email message and the annotated email messages of the database.

62. The article of manufacture of claim 60, wherein the score represents a degree of similarity between the query email message and at least one of the annotated email messages of the database, and wherein said at least one of said annotated email messages comprises an unwelcome email message ("SPAM").

63. The method of claim 60, wherein the score represents a degree of similarity between the query email message and at least one of the annotated email messages of the database, and wherein said at least one of said annotated email messages comprises a welcome email message ("non-SPAM").

64. The article of manufacture of claim 50, further comprising the step of determining a score for the patterns with assigned attributes used to contribute to the attribute vector, said database comprising (i) a first subdatabase comprising annotated unwelcome email messages ("SPAM"), and (ii) a second subdatabase comprising annotated welcome email messages ("non-SPAM"), said score representing a degree of similarity, between the query email message and at least one of said annotated unwelcome email messages ("SPAM"), and a degree of dissimilarity between the query

email message and at least one of said annotated welcome email messages ("non-SPAM").